DOCUMENT RESUME

ED 233 063 TM 830 511

AUTHOR Benson, Jeri; Hocevar, Dennis
TITLE Measuring Scale Invariance between and within Subjects.
PUB DATE Apr 83
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Black Students; *Computer Assisted Testing; *Computer Oriented Programs; Correlation; Elementary Education; Hispanic Americans; Models; *Scaling; *Self Concept Measures; *Test Reliability; White Students
IDENTIFIERS Coopersmith Self Esteem Inventory; *LISREL Computer Program; Scale Analysis; *Test Retest Reliability

ABSTRACT

The present paper represents a demonstration of how LISREL V can be used to investigate scale invariance (1) across time (its relationship to test-retest reliability), and (2) across groups. Five criteria were established to test scale invariance across time and four criteria were established to test scale invariance across groups. Using the Coopersmith Self-Esteem Inventory for Children, six models were developed to test the above criteria with covariance matrices obtained from the responses of 722 Black, White, and Hispanic elementary students. Results indicated that correlated uniquenesses existed across time and this produced an overestimate of the test-retest reliability. In addition, the construct of self-concept was shown to be invariant across the three ethnic groups. Thus, LISREL procedures appear to provide a useful technique for studying scale invariance both within and between subjects. (Author/PN)

MEASURING SCALE INVARIANCE BETWEEN AND WITHIN SUBJECTS

Jeri Benson
Dennis Hocevar
University of Southern California

Abstract

The present paper represented a demonstration of how LISREL V can be used to investigate scale invaraince a) across time and its relationship to test-retest reliability and b) across groups. Five criteria were established to test scale invariance across time and four criteria were established to test scale invariance across groups. Using a well-known self-concept instrument, six models were developed to test the above criteria using covariance matrices obtained from the responses of 722 Black, White and Hispanic elementary students. Results indicated that correlated uniquenesses existed across time and this produced an overestimate of the test-retest reliability. In addition, the construct of self-concept was shown to be invariant across the three ethnic groups. Thus, LISREL procedures appear to provide a useful technique for studying scale invariance both within and between subjects.

Presented at the Annual Meeting of the
American Educational Research Association
Montreal, April 1983

Printed in the United States

Measuring Scale Invariance Between and Within Subjects

The purpose of the present paper was to describe several criteria for evaluating scale invariance and to provide a pedagogical exposition of how scale invariance can be quantified in terms of Joreskog and Sorbom's (1981) LISREL schema. Scales can be invariant in two distinct ways. First, a scale can be invariant across time -- this type of invariance is analogous to test-retest reliability. Second, scales can be invariant across groups -- this type of invariance is similar to the concept of factorial invariance in the factor analytic literature.

The major focus of the present study was invariance across time and its relationship to test-retest reliability. According to Magnusson (1966) reliability can be defined as the 'correlation between two parallel tests' (p 62). Parallel tests can be defined as the same test given on two ocassions or two content-similar tests given on the same ocassion.

Reliability theory is based upon the model presented by Spearman where the observed score for individual j is equal to their true score plus their error score as shown in formula 1.

$$X_j = T_j + E_j . \qquad (1)$$

When different scores result for the same individual based upon the two testings, the difference is attributed to

chance or random error. The assumptions regarding these errors indicates that over an infinite number of testings an individual's mean error score will be zero, the errors are thought not to correlate with the individual's true score and the errors themselves are considered to be uncorrelated (Magnusson, 1966, p.64). Using the above three assumptions, an individual's observed score (X_j) is thought to be a representation of their true score (T_j). Thus, reliability estimates are calculated using the observed test score data and are interpreted as the ratio of true score variance to the total observed test score variance. This interpretation is based upon the above assumptions regarding errors of measurement. Of particular interest here is the assumption that the errors themselves are not correlated for each individual. However, in many testing situations the errors may indeed be correlated. Maxwell (1968) illustrated how correlated errors would effect internal consistency estimates by using an ANOVA model that tested whether the covariance between items was greater than zero. If the item covariances were greater than zero, the internal consistency estimate was considered biased and that the bias could produce an over or an underestimate.

A second focus of present paper was to illustrate how one could investigate scale invariance across independent groups. The invariance of psychometric properties across independent groups has received extensive attention in the factor analytic literature. Typical concerns are the invar-

lance of factor structures, factor variances and covariances, and factor uniquenesses. Prior discussion of invariance across groups using the LISREL procedure have been provided by Benson (1982), Benson, Hocevar and Cohen (1982), Joreskog (1971), McGaw and Joreskog (1971) and Sorbom (1974). In addition, Werts, Rock, Linn and Joreskog (1976) have shown that it is possible to test the equality of variance-covariance matrices between and within subjects with tests of different lengths.

Until recently statistical procedures were not available to test for correlated errors in the test-retest coefficient nor to test simultaneously for scale invariance across groups. With the development of model testing using linear structural relationships (LISREL) developed by Joreskog and Sorbom (1981) the tenability of the assumption of uncorrelated errors of measurement across time can be tested as well as the stability of the scale across groups. Specifically, LISREL V allows the testing of differences in factor structure, true score variance and correlated errors of measurement within and between groups across time. Thus, the major objective of the paper, while using data representing a substantive content area regarding the measurement of self-concept, was mainly a demonstration of how LISREL V can be utilized a) to answer questions regarding the invariance of measurements across time and b) to test scale invariance across groups.

Methodology

Sample

The data represent the scores of elementary students in grades three to six from over 70 schools in a large urban school district. Matched scores were obtained for the students pre to post resulting in a sample of 722. The sample was composed of 395 White students, 213 Black students and 114 Hispanic students; 505 were boys and 217 were girls.

Instrumentation

The instrument used in the study was the Coopersmith Self-Esteem Inventory for Children, Form B. The instrument contains 25 items, eight are positively phrased and 17 are negatively phrased. The response format is in a dichotomous fashion - 'like me' or 'unlike me'. Form B of the scale was developed by Coopersmith (1975) by selecting items which had the highest item/total correlations on Form A, the longer version of the Coopersmith inventory. Due to the nature of the scale's development, the factor structure was assumed to be unidimensional both across groups and within groups. The 25 item scale was administered by an elementary school counselor to the student in both the pre and posttest sessions.

Procedures

The covariance matrix was used as input in testing all models under each of the three scale invariant conditions. In addition, one item was arbitrarily selected as a reference loading and its value was set at 1.0 for all analyses in estimating the Lambda, Phi and Theta values. The Lambda matrix represented the factor loadings (item/total regressions) for each item on the one factor scale. The Phi matrix represented the true score variance for the scale. The Theta matrix represented the item error variance (uniqueness) for each item. Depending upon the model tested, parameters were either set to be invariant (fixed) and given a value of zero or free to be estimated and given a value of one.

Scale Invariance Across Time. Scale invariance across time can be conceptualized in terms of at least five criteria:

1. Are the factor loadings (item-total regressions) invariant across time? This involves simultaneously testing the LISREL item-total regression coefficients (Lambda estimates) from time 1 to time 2.

2. Are the true score variances invariant across time? Statistically, this would involve comparing the estimated true score variance (Phi) for time 1 with the estimated true score variance of time 2.

3. Is the item error variance (uniqueness) invariant across time? This would involve a simultaneous test of the equality of the item error variances (Theta variance estimates) from time 1 to time 2.

4. Are the item uniquenesses for <u>each item</u> at time 1 correlated with their respective item uniquesnesses at time 2? This involves simultaneously testing the correlation of each item's uniqueness at time 1 with it's uniqueness at time 2 (Theta covariance estimates).

5. Are the estimated true scores for time 1 and time 2 correlated? This can be observed by freeing the item uniquesnesses and noting changes in the test-retest reliability estimate.

For studying scale invariance across time, one group of students was arbitrarily selected, the White students. Six models were constructed to test the above questions using LISREL V.

a) Model 1 - The factor structure, true score variance and error variance from time 1 to time 2 were invariant. (Invariant Model)

b) Model 2 - The factor structure was free to vary across time, but the true score variance and error variance from time 1 to time 2 were invariant. (Lambda free)

c) Model 3 The true score variance was free to vary across time, but the factor structure and error variance from time 1 to time 2 were invariant. (Phi free)

e) Model 4 - The total amount of item uniquenesses for each item was free to vary across time, but the factor structure and true score variance from time 1 to time 2 were invariant. (Theta variance free)

d) Model 5 - The individual item error covariances were free to vary across time, but the factor structure and true score variance from time 1 to time 2 were invariant. (Theta covariance free)

f) Model 6 - The factor structure, true score variance and item errors were free to differ from time 1 to time 2. (Unrestricted model)

Scale Invariance Across Groups. Like invariance across time, invariance across groups cannot be assessed by a single criteria. Rather, four related questions can be asked about invariance across groups. For the three groups in the present study the questions were:

1. Are the factor loadings (item-total regressions) invariant across groups? This involves simultaneously testing the LISREL item-total regression coefficients (Lambda estimates) across the three groups.

2. Are the estimated true scores invariant across groups? Statistically, this would involve testing the true score variances (Phi) for equality across the three groups.

3. Is the item uniqueness ariant across groups? This is a test of the invariance of the item uniquenesses (Theta variance estimates) across groups.

4. Is the internal consistency of the scale invariant across groups? This would involve noting the change in the internal consistency (alpha) estimates for each group.

To study scale invariance across groups five models were constructed using the data from all three ethnic groups (Black, White and Hispanic). Models 1, 2, 3, 4 and 6 from above were tested across the three groups.

Results and Discussion

Scale Invariance Across Time

The chi-square tests of model-data fit are reported in Table 1 for the six models tested. The lower the chi-square statistic, the better the model fit the original covariance matrix used as input. All of the chi-square values shown in column 1 were statistically significant at $p<.05$. This finding was in part due to the large sample size. Bentler

and Bonett (1980) have suggested using a chi-square difference test (equation 5) between alternate models to determine the relative effectiveness of model-data fit.

$$\chi^2_{1-2} = \chi^2_1 - \chi^2_2 \quad \text{and} \quad df = df_1 - df_2 \quad , \quad (2)$$

where χ^2_1 represented the most restrictive model and χ^2_2 represented an alternate model with their corresponding degrees of freedom. If the chi-square difference is statistically significant, then the alternate model represents a better fit to the data.

Chi-square difference tests were conducted to answer the first four questions posed in the previous section on testing scale invariance across time. For questions 1-4, the chi-square difference test was conducted by contrasting Model 1 with Models 2-6. The results are shown in Table 1 column 3. For questions 1,2 and 3 the factor structure, true score variance and total item error uniquenesses were found to be invariant across time since the chi- square difference tests were not statistically significant from the invariant model (χ^2_{1-2} = 17.42, df = 24; χ^2_{1-3} = 2.4, df = 1; χ^2_{1-4} = 16.81, df = 25 respectively). However the item error covariances (question 4) were found to be correlated across time (χ^2_{1-5} = 507.14, df = 25, $p<.05$). Thus, the item errors were not independent from time 1 to time 2 and as such, this procedure represented a rejection of the classical test theory assumption regarding uncorrelated errors. In addition, the chi-square difference test between the invariant

model and the unrestricted model was also statistically significant ($\chi^2_{1-6} = 627.22$, df = 75, $p<.05$). This finding meant that the model with correlated error was a better fit to the data than the invariant model which represented a strict definition of classical test theory, where uncorrelated errors were assumed. For this set of data then, the assumption of uncorrelated errors across testings did not appear tenable and it was basically this difference that resulted in the unrestricted model being a better fit to the data than the invariant model.

Finally, to emphasize the improvement in model-data fit of Models 5 and 6 over Model 1, the delta index (Bentler & Bonett, 1980) was calculated and is shown in Table 1 column 4. Delta represents an incremental index of fit that is independent of sample size and is calculated as

$$\Delta_{12} = \chi^2_1 - \chi^2_2 \; / \; \chi^2_1 \quad , \qquad (3)$$

where χ^2_1 is thought to represent the most restrictive model and χ^2_2 is an alternative model. The values of delta range between zero and one. The results parallel that of the chi-square difference test, where Models 5 and 6 represent a better fit of the original covariance matrix than Model 1 (.209 and .258, respectively).

Question 5, regarding the possible bias in test-retest reliability due to correlated errors, was tested by noting the difference in the phi matrix from Model 1 to Models 5 and 6. The off-diagonal of the phi matrix gives the amount

of covariance between the true variance of time 1 and time 2. This value, adjusted by the standard deviations of the true variances for time 1 and time 2, represents the correlation or test-retest reliability. Under the condition of complete invariance (Model 1), the test-retest coefficient was .630. When the errors were allowed to be correlated (Models 5 and 6) the test-retest coefficient was .600 and .570 respectively. Therefore, when measurement errors are correlated between testings, the test-retest reliability may be over or underestimated. For this set of data, the overestimate was very slight however, it may not be so with other data. Thus, psychometricians can test for correlated errors and adjust for them, if need be, by using LISREL procedures.

Scale Invariance Across Groups

The chi-square statistic for model-data fit is reported in Table 1 column 1 for the five models tested. All chi-square values were statistically significant at $p<.05$. To answer the first four questions posed for scale invariance across groups, the chi-square difference test was run comparing Model 1 to Models 2-5. The results are shown in Table 1 column 3.

For Questions 1 and 3, the factor structure and item uniquenesses were found to be invariant across the three ethnic groups ($\chi^2_{1-2} = 64.57$, df = 48; $\chi^2_{1-4} = 42.35$, df= 50, $p>.05$, respectively). The delta index of incremental fit

for Models 2 and 4 was also very small (.041 and .027, respectively). Regarding Question 2, the true score variance was found to differ significantly across the three ethnic groups (χ^2_{1-3} = 9.1, df = 2, $p<.05$). However, the delta index of fit indicated that this difference was not practically significant (.006) and illustrated how large sample sizes can highlight trivial differences by using only the chi-square test or the chi-square difference test. Also, the unrestricted model was not superior to the invariant model using the chi-square difference test due to the large difference degrees of freedom (χ^2_{1-5} = 111.04, df = 100, $p>.05$). Although slight, the delta index was greatest for the unrestricted model (.071) again, indicating no practical significance from the strict invariant model. Thus, the factor structure, the amount of true score variance and the item uniquesness were invariant across the groups. This procedure allows one to test the difference, if any, in the construct being measured for each group. For this set of data, the construct being measured was shown to be invariant across groups and represented a test of factorial stability.

For question 4, regarding the invariance of the scale's internal consistency across groups, the alpha reliability coefficient for time 1 for the White group was .78, for the Black group .70 and for the Hispanics .60. Overall all groups, the reliability was .74 for time 1. For time 2, the reliability coefficient for the White group was .81, for the Black group .74 and for the Hispanics .74. The alpha relia-

bility for all groups was .79 for time 2. A slight increase from time 1 to time 2 was noted for the White and Black groups and a rather large increase was observed for the Hispanic group. Since it was shown earlier that the item uniquenesses were correlated for the White group from time 1 to time 2, it may be that the item uniquenesses within each group may likewise be correlated. A test for correlated errors has been reported by Maxwell (1968) and could be used to determined if the differences noted in the internal consistency estimates above were true differences or were due to correlated errors within each group which could produced an over or underestimate of the internal consistency at time 2.

Conclusions

An approach to testing scale invariance across time and groups was demonstrated using LISREL V. The importance of testing for scale invariance across time is that the construct being measured may vary from time 1 to time 2 in terms of it's factor structure, true score and error variance as well as the accurracy or the stability of the measurement. If the the construct being measured changes from time 1 to time 2, then problems in interpretation of the construct will occur. If the item uniquenesses are correlated from time 1 to time 2, then the test-retest reliability coeficient will be biased. As was shown in the present study, the factor structure and amount of true score and er-

ror variance did not change across time, but the item uniquenesses were correlated from time 1 to time 2 and resulted in an overestimate of the test-retest coefficient.

Secondly, it was demonstrated that one can test for scale invariance across groups. This is a test of the stability of the construct being measured for the groups involved. If the construct varies across independent groups then the confidence one would have in the interpretation made of the observed score would not be very strong. Testing for scale invariance across groups using LISREL provides a way to confirm or disconfirm the similarity of the construct being measured for each group. For the present study, the unideminsional construct of self- concept was invariant across the three ethnic groups studied, although the degree of item homogeneity within groups differed. The difference in item homogeneity may be attributed to correlated errors within each group and could be tested using an analysis of variance model proposed by Maxwell (1968). LISREL procedures are potentially very useful as they allow for the testing of correlated errors and their effect on scale invariance across time and the testing of scale invariance across groups.

REFERENCES

Benson, J. Detecting bias in affective scales. Paper presented at the annual meeting of the American Educational Association, New York, 1982.

Benson, J., Hocevar, D. & Cohen, C. Effects of item phrasing on the factorial invariance of attitude measures in elementary school children. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982.

Bentler, P.M. & Bonett, D.G. Significant tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 1980, 88, 588-606.

Coopersmith, S. Coopersmith Self-Esteem Inventory, Technical Manual. Palo Alto, Ca.: Consulting Psychologists Press, Inc., 1975.

Joreskog, K.G. Simultaneous factor analysis in several populations. Psychometrika, 1971 36, 409-426.

Joreskog, K.G. & Sorbom, D. LISREL V: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood. Chicago, Ill.: International Educational Services, 1981.

McGaw, B. & Joreskog, K.G. Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. British Journal of Mathematical and Statistical Psychology, 1971, 24, 154-168.

Magnusson, D. Test Theory. Reading, Mass.: Addison-Wesley, 1966.

Maxwell, A.E. The effect of correlated errors on estimates of reliability coefficients. Educational and Psychological Measurement, 1968, 28, 803-811.

Sorbom, D. A general method for studying differences in factor means and factor structures between groups. British Journal of Mathematical and Statistical PSychology, 1974, 27, 229-239.

Werts, C.E., Rock, D.A., Linn, R.L. & Joreskog, K.G. A comparison of correlations, variances, covariances and regression weights with or without measurement errors. Psychological Bulletin, 1976, 83, 1007-1013.

Table 1

Goodness of Fit Indices for Models Tested

	Chi-Square	df	χ^2_{diff}	Delta
Scale Invariance Across Time				
1. All Invariant	2427.47	1224	------	----
2. Lambda free	2410.05	1200	17.42/24	.007
3. Phi free	2425.07	1223	2.40/1	.000
4. Theta free (Variance)	2410.66	1199	16.81/25	.007
5. Theta free (Covariance)	1920.33	1199	507.14/25*	.209
6. All free	1800.25	1149	627.22/75*	.258
Scale Invariance Across Groups				
1. All Invariant	1558.74	925	------	----
2. Lambda free	1494.17	877	64.57/48	.041
3. Phi free	1549.60	923	9.14/2 *	.006
4. Theta free (Variance)	1516.39	875	42.35/50	.027
5. All free	1447.70	825	111.04/100	.071

*$p<.05$